The Annals of Applied Statistics
2011, Vol. 5, No. 2B, 1553–1585
DOI: 10.1214/10-AOAS447

A HIERARCHICAL BAYESIAN APPROACH TO RECORD LINKAGE AND POPULATION SIZE PROBLEMS[1]

BY ANDREA TANCREDI AND BRUNERO LISEO

Sapienza Università di Roma

We propose and illustrate a hierarchical Bayesian approach for matching statistical records observed on different occasions. We show how this model can be profitably adopted both in record linkage problems and in capture–recapture setups, where the size of a finite population is the real object of interest. There are at least two important differences between the proposed model-based approach and the current practice in record linkage. First, the statistical model is built up on the actually observed categorical variables and no reduction (to 0–1 comparisons) of the available information takes place. Second, the hierarchical structure of the model allows a two-way propagation of the uncertainty between the parameter estimation step and the matching procedure so that no plug-in estimates are used and the correct uncertainty is accounted for both in estimating the population size and in performing the record linkage. We illustrate and motivate our proposal through a real data example and simulations.

1. Introduction. The current explosion in the availability of data from multiple sources, and the relative ease of information storage have led to a great popularity of statistical methods which aim at merging and/or matching statistical information available from different sources. Among these methods, record linkage refers to the problem of identifying statistical units which may be present in more than one data set. Fienberg and Manrique-Vallier (2009) review the relevance of record linkage procedures in official statistics and highlight the significant intertwins with missing data and multiple systems estimation literature.

The gist of this paper is the proposal of a hierarchical Bayesian framework which can be profitably adopted both in record linkage problems and in capture–recapture scenarios, where the size of a finite population is the main object of interest and the number of "re-captured" individuals is unknown. Most of the current approaches to population size estimation with matching uncertainty consider the matching and the size estimation as two logically well separated steps. Remarkable recent exceptions are Link et al. (2009) and Wright et al. (2009) where genotype misidentification is embedded into multiple mark-recapture models for estimating animal abundance using DNA samples. More generally, in this paper, we propose

Received May 2009; revised November 2010.

[1]Supported by MIUR Grant 2008 CEFF37;002, Bayesian Methods for finite populations, Italy.

Key words and phrases. Capture–recapture methods, conditional independence, Gibbs sampling, Metropolis–Hastings, record linkage.

a unified framework where matching uncertainty is naturally accounted for in estimating population size by using samples of multivariate categorical variables.

To motivate our approach, consider the following example, which is a part of a real application. Suppose we have two data sets which we call A and B with sizes, respectively, 34 and 45. Data set A comprises all the foreign residents observed in a small census block during the 2001 Italian census population survey (CPS). Data set B comprises all the resident foreigners observed in the same census block during the post enumeration survey (PES)[2]. Both data sets report, among others, the following variables: (1) first two consonants of the family name, (2) gender and (3) education level. Assume that the three variables represent the only available information to perform the match; assume also that the goal is the estimation of N, the total number of foreign residents in the census block. The usual approach to this problem would be to search for the pairs of units, belonging to different files, which agree perfectly on each observed variable. In our example there are 25 pairs which show a complete agreement. If we assume that we actually observed 25 recaptures, such information can be used easily in a capture–recapture model to make inference on N. However, two complications may arise. First, it might be possible that two different units genuinely agree on each variable. Second, because of measurement error, observed records for the same unit might be different in the two sampling occasions. They could also agree as before, even if they refer to different units with different true values. We will discuss this example below in more detail. For the moment, Table 1 summarizes, for the different choices of the declared number T of recaptures, the posterior distribution of N assuming a noninformative prior $p(N) \propto 1/N^2$ and a hypergeometric likelihood function $p(T|N) \propto \binom{n^A}{T}\binom{N-n^A}{n^B-T}/\binom{N}{n^B}$ with $n^A = 34$ and $n^B = 45$. One can see that slightly different choices of T may produce dramatically different posterior distributions.

Accounting for matching uncertainty has relevance well beyond size estimation problems and relates to the more general problem of inference with integrated data;

TABLE 1

Posterior quantiles for N with the distribution $p(N|T) \propto \binom{n^A}{T}\binom{N-n^A}{n^B-T}/\binom{N}{n^B} \times (1/N^2)$ with $n^A = 34$, $n^B = 45$ and different choices of T

	T						
	24	**25**	**26**	**27**	**28**	**29**	**30**
2.5%	57	56	54	53	51	50	49
50%	64	62	59	57	55	53	51
97.5%	78	74	70	66	63	60	57

[2]PES is usually performed some time after CPS, to evaluate the effective coverage of CPS.

see Judson (2007). In this context, an important exemplification is provided by Lahiri and Larsen (2005) who take into account linkage uncertainty in the framework of the linear regression model when the response variable and the covariates are recorded on two different occasions. However, our approach can also be applied when there is not yet a scheduled statistical analysis to be performed on the linked data, but the linkage procedure is just the initial step to obtain a larger and integrated reference data set.

Statistical methods for finding entries related to the same entity in two or more files are employed in many different disciplines, such as medicine, business administration and official statistics [see, e.g., Herzog, Scheuren and Winkler (2007)]. In these contexts it may happen that a unique data set with all the necessary information for a particular statistical analysis is not available. Furthermore, time and cost constraints may make it unfeasible to create such a data set anew. Integration at the unit level of different data sets (sample surveys and/or administrative data sets) may be an answer to this kind of problem. A considerable difficulty in this context is represented by the lack of a unique identifier in the different data sets for each unit of interest. In fact, when a set of observed variables (key variables, henceforth) may be used as an identifier for connecting records that refer to the same unit, particular attention should be paid to errors, as we have seen in the introductory example, and missing values.

To handle the record linkage process, many different methodologies have been introduced. Some methods are naïve, or heuristic, that is, are based only on common sense [e.g., the "iterative method" described in Armstrong and Mayda (1993)]. In a fundamental paper, Fellegi and Sunter (1969) put these kinds of problems into a firm, model based, statistical framework. Further advances were described in a number of papers in the 1980s and 1990s: among others, Jaro (1989), Winkler (1993) and Belin and Rubin (1995). Larsen and Rubin (2001) introduce the representation of the record linkage problem in terms of the mixture model [see also Larsen (1999)]: this idea has been exploited in many other papers; see, for example, Fortini et al. (2001), McGlincy (2004) and Larsen (2004) who tackle the problem from a Bayesian perspective. All of these papers assume that each single comparison between records in two different files provides new information, independently of the other comparisons. This assumption, as noted by Kelley (1986), is fundamentally unsound, as illustrated in Section 2. Also, in this respect, Winkler (2000) states that *"... because the underlying true probabilities have not been accurately estimated, estimated error rates (of the record linkage procedure) are not accurate."*

An important feature of our paper is that we propose a Bayesian model which is based on the actual observed data rather than comparisons. In a similar spirit, Fortini et al. (2002) discussed these ideas in the simple setting of a single continuous variable. Here we will assume that our key variables will be discrete, as almost always happens in practice.

Record linkage is not the only statistical problem where matching issues are concerned. In a bioinformatics context Green and Mardia (2006) introduce a matching matrix (very similar to our matrix C, see later) into some problems of shape analysis, where configurations of points in space need to be matched and the points are not completely labeled.

DeGroot and Goel (1980) consider the situation where a random sample of size n, say, (X_i, Z_i), $i = 1, \ldots, n$, is drawn from a bivariate normal distribution; however, before the sample values are recorded, each observation $(x_i; z_i)$ gets broken into two separate components. As a consequence, the available information is represented by the vectors $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$, where y is an unknown permutation of the values $(z_1, \ldots, z_n)$.

Another matching example is discussed in Lindley (1977), in a forensic framework. Here the matching problem arises when some material is found at the scene of a crime and similar material is found on a suspect; in both cases material collection is subject to measurement error. Lindley describes a Bayesian method to establish whether the two materials come from the same source or not. When rephrasing Lindley's approach from a record linkage perspective, we note that that paper was the first attempt to introduce, into a Bayesian linking model, the natural idea that two units with the same surname are more likely to be a match if the surname is *Bodolomonogoto* than if the surname is *Smith*. Similar suggestions can be found in the seminal papers by Newcombe et al. (1959) and Fellegi and Sunter (1969).

The paper is structured as follows. In Section 2 we present the standard approach to record linkage. Our Bayesian approach is discussed in Section 3. Markov chain Monte Carlo (MCMC) methods are needed for estimating the parameters of the model. In Section 4 we describe a suitable algorithm for simulating the posterior distribution. We also discuss a loss function approach to the matching estimation. In Section 5 the performance of the methodology is evaluated through a small illustrative application. A more realistic example is shown in Section 6. A simulation study is conducted in Section 7. Finally, in Section 8 we give a brief discussion of possible future extensions and improvements of the method.

2. Classic approach to record linkage. Suppose we are given two record configurations x^A and x^B of different sizes n^A and n^B with

$$x^A = (x_1^A, \ldots, x_a^A, \ldots, x_{n^A}^A)' \quad \text{and} \quad x^B = (x_1^B, \ldots, x_b^B, \ldots, x_{n^B}^B)'.$$

Here $x_a^A = (x_a^{A_1}, \ldots, x_a^{A_h})$ and $x_b^B = (x_b^{B_1}, \ldots, x_b^{B_h})$ are the observed values of a categorical random vector $x = (x^1, \ldots, x^h)$ whose support is the set

$$V = \{v_{j_1 j_2, \ldots, j_h} = (v_{j_1}^1, v_{j_2}^2, \ldots, v_{j_h}^h),\ j_1 = 1, \ldots, k_1; \ldots; j_h = 1, \ldots, k_h\}.$$

In the following, the two data configurations will be called, respectively, sample A and sample B, the components of the random vectors x (whenever it is possible we will avoid subscript and superscript indices to simplify the notation) are the

key variables and the elements of the set V arranged in lexicographic order will be indicated with v_j for $j = 1, \ldots, k = k_1 \cdot k_2 \cdots k_h$.

Let $A \times B$ be the set of all possible pairs of units belonging to different samples. Set $A \times B = M \cup U$, where $M = \{(a, b) \in A \times B : a \equiv b\}$ (here $a \equiv b$ means that unit a of sample A and unit b of sample B are the same population unit) and $U = \{(a, b) \in A \times B : a \not\equiv b\}$. Probabilistic record linkage, as implemented, for example, in Jaro (1989), is performed by modeling the comparison vectors $y_{ab} = (y1_{ab}, \ldots, y^h_{ab})$ where

$$y^i_{ab} = \begin{cases} 1, & x_a^{A_i} = x_b^{B_i}, \\ 0, & x_a^{A_i} \neq x_b^{B_i}, \end{cases} \qquad i = 1, \ldots, h.$$

Vectors y_{ab}, $a = 1, \ldots, n^A$, $b = 1, \ldots, n^B$, are assumed independent conditionally on M and U. The probability distribution of y_{ab} depends on the match or nonmatch status of the single pair (a, b); in particular, it is assumed that $p(y_{ab}|(a, b) \in M) = \prod_{i=1}^h m_i^{y^i_{ab}}(1 - m_i)^{1-y^i_{ab}}$ and $p(y_{ab}|(a, b) \in U) = \prod_{i=1}^h u_i^{y^i_{ab}}(1 - u_i)^{1-y^i_{ab}}$ (here and later, we will abuse notation by letting the arguments define the functions) with $m = (m_1, \ldots, m_h)$ and $u = (u_1, \ldots, u_h)$ as unknown probabilities vectors. In addition, the elements of the sets M and U are modeled assuming that each pair in $A \times B$ is a match with probability w, independently of all the other pairs. This way the comparison vectors y_{ab} are independent and identically distributed as a mixture of two multivariate Bernoulli distributions:

$$\begin{aligned} p(y_{ab}|m, u, w) = w \prod_{i=1}^h m_i^{y^i_{ab}}(1 - m_i)^{1-y^i_{ab}} \\ + (1 - w) \prod_{i=1}^h u_i^{y^i_{ab}}(1 - u_i)^{1-y^i_{ab}}. \end{aligned} \tag{2.1}$$

Models similar to (2.1) are often used also in biostatistics, under the name of a latent class model, to assess diagnostic test accuracy in the absence of a gold standard and only multiple imperfect tests are available [Pepe (2003)]. Likelihood maximization of the parameters in model (2.1) is performed via the EM algorithm and analytical expressions for the estimators are provided by Fellegi and Sunter (1969) and Pepe and Janes (2007) in the case $h = 3$.

Several extensions of this basic setup have been proposed; see, for example, Larsen and Rubin (2001). In order to decide whether to declare a link a single pair, one can consider the likelihood ratio

$$\lambda = \frac{P(y_{ab}|(a, b) \in M)}{P(y_{ab}|(a, b) \in U)} = \frac{\prod_{i=1}^h m_i^{y^i_{ab}}(1 - m_i)^{1-y^i_{ab}}}{\prod_{i=1}^h u_i^{y^i_{ab}}(1 - u_i)^{1-y^i_{ab}}} \tag{2.2}$$

or the posterior probability

$$p\big((a,b)\in M|y_{ab}\big) = \frac{w\prod_{i=1}^{h} m_i^{y_{ab}^i}(1-m_i)^{1-y_{ab}^i}}{w\prod_{i=1}^{h} m_i^{y_{ab}^i}(1-m_i)^{1-y_{ab}^i} + (1-w)\prod_{i=1}^{h} u_i^{y_{ab}^i}(1-u_i)^{1-y_{ab}^i}}. \tag{2.3}$$

Pairs with high values of λ or $p((a,b)\in M|y_{ab})$ are then declared matches. This approach is formalized in the classical approach of Fellegi and Sunter (1969).

In our opinion the above approach can be criticized on several grounds:

1. *Decision rules for classifying records as matches.* In general, all the pairs with a likelihood ratio λ, or a posterior probability, above a fixed threshold are declared matches. However, the choice of the threshold can be problematic, as illustrated, for example, in Belin and Rubin (1995). More details about this point will be given in Section 4.
2. *Avoiding multiple matches.* Current approaches to record linkage assume that there are no duplications in the same file and inference procedures should account for that. However, in classical procedures, it might happen that a single record in A is linked to more than one record in B; consequently, some extra assumptions are necessary. Jaro (1989) proposes a linear programming approach after a preliminary match estimation step. An alternative approach [Fortini et al. (2001)], which will be pursued here, incorporates the constraints into the sampling model.
3. *Incorporating sampling information.* If we assume that the two files are random samples without replacement from a population of unknown size N, an obvious prior assumption is $p((a,b)\in M) = 1/N$, with $N > \max\{n^A, n^B\}$. In addition, if we know that two units assume the same value v_j, the matching probability becomes $p((a,b)\in M) = 1/F_j$, where F_j is the (unknown) total number of units with record v_j in the population. In record linkage procedures, in general, sources of knowledge of this type are not included in the model, with an obvious loss of information. This may be particularly important for applications of record linkage in disclosure literature.
4. *Comparison vectors are not independent.* In this respect Kelley (1986) states: *"... The decision procedure ... was developed under the hypothesis that the comparison vectors between separate record pairs are independent. However, since the record pairs that are considered for possible matches are elements of the cross product of the two files we are attempting to match, the comparison vectors are in fact dependent ...."* As a matter of fact, the random variables y_{ab} are deterministically dependent. To see that, consider the case of one key variable X_1. Suppose that $x_1^{A_1} = x_1^{B_1}$ and $x_1^{A_1} = x_2^{B_1}$. If, in addition, $x_2^{A_1} = x_1^{B_1}$, it must necessarily be true that $x_2^{A_1} = x_2^{B_1}$, that is, in terms of comparisons, $p(y_{22} = 1|y_{11} = 1, y_{12} = 1, y_{21} = 1) = 1$. Moreover, the problem of

dependency among the y_{ab}'s cannot be circumvented by eliminating redundant comparisons in the likelihood function, because the order in which pairs are considered would matter!

5. *The components y^i_{ab} of the comparison vector may not be independent conditionally on M and U.* The conditional independence assumption among the key variables often fails in practice: disagreement on different key variables for a true match might be caused by a unique reason which introduces correlation among the y^i_{ab}'s. In the absence of conditional independence, the resulting estimates of w, m and u lose their meaning and a more sophisticated conditional dependence structure must be specified. Similar arguments have been applied to criticize the use of model (2.1) for the analysis of diagnostic test performance without a gold standard and, in this context, several solutions have been proposed and discussed [Albert and Dood (2004); Pepe and Janes (2007)]. Larsen and Rubin (2001) have introduced interactions among key variables; see also Winkler (1995) and references therein.

3. The new model. We assume that the records in x^A and x^B are measurements subject to recording error of a multivariate categorical variable $\mu = (\mu 1, \ldots, \mu^h)$ whose support is, on both occasions, the set V. Specifically, let

$$\mu^A = (\mu^A_1, \ldots, \mu^A_a, \ldots, \mu^A_{n^A})' \quad \text{and} \quad \mu^B = (\mu^B_1, \ldots, \mu^B_b, \ldots, \mu^B_{n^B})'$$

be two independent random samples from the multivariate categorical variable μ drawn on different occasions from the same finite population. Let $\mu^A_a = (\mu^{A_1}_a, \ldots, \mu^{A_h}_a), a = 1, \ldots, n^A$ and $\mu^B_b = (\mu^{B_1}_b, \ldots, \mu^{B_h}_b), b = 1, \ldots, n^B$ be the unobserved true values for unit a in sample A and unit b in sample B. We assume that, conditionally on their respective true values and a parameter vector $\beta = (\beta_1, \ldots, \beta_h)$ which accounts for the measurement error, x^A and x^B are independent, that is,

$$p(x^A, x^B | \mu^A, \mu^B, \beta) = p(x^A | \mu^A, \beta) p(x^B | \mu^B, \beta);$$

we also assume that, in each sample, all the observations are conditionally independent given their true values and β. Then

$$p(x^A | \mu^A, \beta) = \prod_{a=1}^{n_A} p(x^A_a | \mu^A_a, \beta), \qquad p(x^B | \mu^B, \beta) = \prod_{b=1}^{n^B} p(x^B_b | \mu^B_b, \beta),$$

with

$$p(x^A_a | \mu^A_a, \beta) = \prod_{i=1}^{h} p(x^{A_i}_a | \mu^{A_i}_a, \beta_i), \qquad p(x^B_b | \mu^B_b, \beta) = \prod_{i=1}^{h} p(x^{B_i}_b | \mu^{B_i}_b, \beta_i).$$

Note that the vectors μ^A and μ^B introduce a first latent structure into our record linkage model and make it effectively a missing data model [Fienberg and Manrique-Vallier (2009)].

We conclude the top stage of the hierarchical structure by explicitly introducing the measurement error model. A general model for potentially misclassified observed records can be formulated as $p(x^i = v^i_{j_i} | \mu^i = v^i_{j'_i})$, for all (j_i, j'_i). Such a model has been considered, in a Bayesian framework, by Swartz et al. (2004) who discuss several identifiability problems, and by Perez et al. (2007), where strong prior information is introduced in the model. Here, to maintain the number of parameters in the model reasonably low, we propose a simpler version of the so-called *hit–miss* model [Copas and Hilton (1990)]

$$p(x^i = v^i_{j_i} | \mu^i = v^i_{j'_i}) = \beta_i I(v^i_{j_i} = v^i_{j'_i}) + (1 - \beta_i)/k_i, \qquad i = 1, \ldots, h,$$

where β_i represents the probability of observing the true value for the ith variable "not by chance" and k_i is the number of levels of variable x^i. This way, conditionally on the unobserved true values, each single record field can be modeled as a mixture of two components: the first component is concentrated on the true value, while the second one is uniformly distributed over the set $v^i = \{v^i_1, \ldots, v^i_{k_i}\}$. For a recent implementation of the hit–miss model see also Norén, Orre and Bate (2005).

We now specify the conditional distributions of μ^A and μ^B. In particular, we assume that μ^A and μ^B are two independent simple random samples drawn without replacement from a finite population of unknown size N. The unknown vector $F = (F_1, \ldots, F_j, \ldots, F_k)$, $k = \prod_{i=1}^h k_i$, represents the population counts for each element v_j of the set V. Obviously, $\sum_{j=1}^k F_j = N$. In principle, one can write the model for the unobserved true values μ^A and μ^B in the following natural way:

$$p(\mu^A, \mu^B | F) = p(\mu^A | F) p(\mu^B | F) \tag{3.1}$$

with

$$p(\mu^S | F) = \binom{n^S}{f^S_1, \ldots, f^S_k}^{-1} \left[\binom{N}{n^S}^{-1} \prod_{j=1}^k \binom{F_j}{f^S_j} \right] \qquad S = A, B, \tag{3.2}$$

where $f^S = (f^S_1, \ldots, f^S_j, \ldots, f^S_k)$, $S = A, B$, are the true sample counts (which are, however, unobservable, due to measurement error) for each element $v_j \in V$. Formula (3.2) can be obtained by noticing that the observed values of μ^S determine the frequencies f^S, so that $p(\mu^S | F) = p(\mu^S | f^S, F) p(f^S | F)$ where $p(\mu^S | f^S, F)$ and $p(f^S | F)$ correspond to the two terms in (3.2). The usual constraints $0 \le f^S_j \le F_j$, $S = A, B$, must hold.

An alternative way of writing the above model is based on the use of two latent quantities, which will play a crucial role in our approach. The first quantity is the so-called matching matrix C. This is a $n^A \times n^B$ matrix whose generic element C_{ab} is a Bernoulli random variable indicating whether or not unit a in sample A and unit b in sample B are the same unit, that is,

$$C_{ab} = \begin{cases} 1, & \text{if } (a, b) \in M, \\ 0, & \text{if } (a, b) \in U. \end{cases}$$

The matrix C is the actual quantity of interest in record linkage problems; a similar structure also appears in different statistical problems, such as the Bayesian alignment [Green and Mardia (2006)] or microarrays analysis [Do, Mueller and Tang (2005)]. We assume that multiple matches are not possible. This implies that $\sum_a C_{ab} \leq 1 \ \forall b = 1, \ldots, n^B$, $\sum_b C_{ab} \leq 1 \ \forall a = 1, \ldots, n^A$; also, note that there are $\binom{n^A}{T}\binom{n^B}{T}T!$ different C matrices with exactly $T = \sum_{ab} C_{ab}$ matches, $T \leq \min(n^A, n^B)$.

The other latent quantity we introduce is the vector $t = (t_1, \ldots, t_j, \ldots, t_k)$ denoting, for each element of V, the number of matches having v_j as the true value. The vector t (which is basically needed to facilitate the simulation of the posterior distribution, as outlined in the following section) is a deterministic function of μ^A, μ^B and C.

Consider, as an illustration, the case where μ is univariate and $V = \{v_1, v_2, v_3, v_4\}$: suppose we have $\mu^A = (v_1, v_2, v_1)$, $\mu^B = (v_2, v_3, v_1, v_2)$, with $C_{13} = C_{24} = 1$ and all the other elements of C equal to 0; then $t = (1, 1, 0, 0)$. Finally, notice that $0 \leq t_j \leq \min\{f_j^A, f_j^B\} \ \forall j = 1, \ldots, k$ and $\sum_{j=1}^k t_j = T$.

Now we introduce the model assumptions for the conditional distribution of μ^A and μ^B given the values of t, C and F. First, note that $p(\mu^A, \mu^B | t, F, C) = 0$ when $\mu_a^A \neq \mu_b^B$ and $C_{ab} = 1$. Also, we have $p(\mu^A, \mu^B | t, F, C) = 0$ either when $\min\{f_j^A, f_j^B\} < t_j$ or $\max\{f_j^A, f_j^B\} > F_j$. In any other situation it turns out that

$$
\begin{aligned}
p(\mu^A, \mu^B | C, t, F) = {} & \frac{\prod_{j=1}^k \binom{F_j - t_j}{f_j^A - t_j, f_j^B - t_j, F_j - f_j^A - f_j^B + t_j}}{\binom{N-T}{n^A - T, n^B - T, N - n^A - n^B + T}} \\
& \times \frac{\prod_{j=1}^k t_j!(f_j^A - t_j)!(f_j^B - t_j)!}{T!(n^A - T)!(n^B - T)!}.
\end{aligned} \tag{3.3}
$$

The distribution in (3.3) has the following interpretation: the first term is the joint distribution of the sample counts f^A and f^B, say, $p(f^A, f^B | C, t, F)$; it can be obtained by observing that, given the vector t, there are already t_j elements in the category v_j, $j = 1, \ldots, k$. Then, out of the total number of partitions of the $N - T$ elements actually sampled in three disjoint sets[3] of sizes $n^A - T$, $n^B - T$ and $N - n^A - n^B + T$, one should only consider those where category v_j respectively appears $f_j^A - t_j$, $f_j^B - t_j$ and $F_j - f_j^A - f_j^B + t_j$ times in the three sets, for $j = 1, \ldots, k$. The other term in (3.3) is the conditional distribution $p(\mu^A, \mu^B | f^A, f^B, C, t, F)$; given f^A and f^B, the matching matrix C and the vector t, there are

$$
T!(n^A - T)!(n^B - T)!
$$

[3]They respectively represent the "nonmatch" for samples A and B and the "nonsampled" units.

possible permutations of the elements of the two samples: among them, there are $\prod_j t_j!(f_j^A - t_j)!(f_j^B - t_j)!$ permutations which exactly reproduce the orderings given in μ^A and μ^B.

The prior distribution for C and t should reflect the random selection mechanism of the two samples. Conditionally on t and F, C has a uniform distribution on the set of all possible matching matrices with T matches. Loosely speaking, in the absence of information about μ^A and μ^B, all the possible couples are equally likely to be a match. Then, we have $p(C, t|F) = p(C|t, F)p(t|F)$ with

$$p(C|t, F) = p(C|t) = \begin{cases} 0, & \text{if } \sum_{ab} C_{ab} \neq \sum_j t_j, \\ \left[\binom{n^A}{T}\binom{n^B}{T} T!\right]^{-1}, & \text{otherwise.} \end{cases}$$

To derive the distribution $p(t|F)$, one can observe that, given T, t is a vector of counts of the v_j categories in a simple random sample of size T drawn from the population. Then $p(t|T, F)$ is a multivariate hypergeometric distribution. Finally, T, the total number of common units across the two samples, is a scalar hypergeometric random variable. Then,

$$\begin{aligned} p(t|F) &= p(t|T, F)p(T|F) \\ &= \prod_{j=1}^{k} \left[\binom{F_j}{t_j} \Big/ \binom{N}{T}\right] \binom{n^A}{T}\binom{N - n^A}{n^B - T} \Big/ \binom{N}{n^B}. \end{aligned} \tag{3.4}$$

It is easy to see that, by averaging out over C and t in the distribution $p(\mu^A, \mu^B, C, t|F)$, one re-obtains the model expressed by (3.1) and (3.2). Details are given in Appendix A. For the moment notice that the use of the hypergeometric distribution $p(T|F)$ in (3.4) is standard practice in capture–recapture modeling when the number T of common units across two samples is known [Darroch (1958), Seber (1986) and Marin and Robert (2007)].

At the bottom of the hierarchical model, one needs to specify the prior for the vector F; this is equivalent to assuming that the finite population which the two samples are drawn from is itself a random sample from a superpopulation model [Ericson (1969)]. In particular, following Hoadley (1969), we assume that, conditionally on N and a vector $\theta = (\theta_1, \ldots, \theta_k)$, with $0 \leq \theta_i \leq 1$ and $\sum_{i=1}^k \theta_i = 1$, F is a multinomial random variable,

$$p(F_1, \ldots, F_k|\theta, N) = \frac{N!}{F_1!F_2! \cdots F_k!} \prod_{j=1}^{k} \theta_j^{F_j}.$$

Regarding the prior for N, we suggest the following family of noninformative priors:

$$p_g(N) \propto \Gamma(N - g + 1)/N!, \qquad g \geq 0;$$

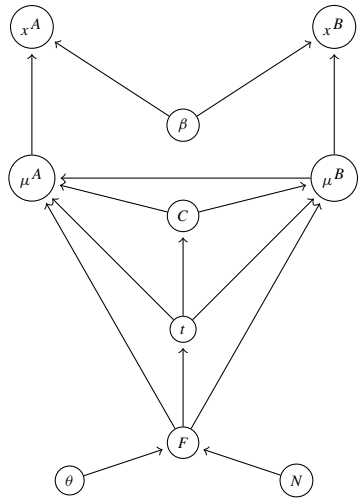

FIG. 1. *DAG representation of the joint probability model described in Section 3.*

the hyperparameter g regulates the shape of the prior: the larger the value of g, the lower the prior weight on the right tail, which is integrable for all $g > 1$. The same prior model for F can be expressed by assuming that, for a fixed hyperparameter $\lambda > 0$ and θ, the population counts $F_1, \ldots, F_k$ are independent Poisson variables with rates $\lambda\theta_1, \ldots, \lambda\theta_k$ and $p_g(\lambda) \propto 1/\lambda^g$.

Last, we assume that the prior for θ is obtained first by modeling its elements via the product of marginal and conditional probabilities based on a specific association pattern for the key variables and then by putting independent Dirichlet distributions to each probability vector characterizing the resulting model for θ. A special case of this product of Dirichlet distributions is the hyper-Dirichlet prior which is used in the similar context of disclosure risk assessment by Forster and Webb (2007); see also O'Hagan and Forster (2004). Moreover, the "measurement error" parameters β are independent and uniformly distributed random variables; they are also independent of all the other model parameters. To sum up, the joint distribution of all the variables is expressed by the following factorization:

$$\begin{aligned} p(x^A, x^B, \mu^A, \mu^B, \beta, C, t, F, N, \theta) = {} & p(x^A, x^B | \mu^A, \mu^B, \beta) \\ & \times p(\mu^A, \mu^B | F, C, t) p(C|t) p(t|F) \\ & \times p(F|\theta, N) p(N) p(\theta) p(\beta), \end{aligned}$$

and a representation in terms of a directed acyclic graph is displayed in Figure 1.

4. Bayesian implementation. In this section we discuss a *Metropolis within Gibbs* algorithm for simulating from the joint posterior distribution $p(\mu^A, \mu^B, \beta, t, F, N, \theta \mid x^A, x^B)$; see Robert and Casella (2004) for a general overview about

the MCMC theory and implementation. In our case the Gibbs algorithm structure is based on the following updating steps:

$$\begin{aligned}&\mu^A, \mu^B, t | F, N, \theta, \beta,\\ &\quad F, N | \mu^A, \mu^B, t, \theta, \beta,\\ &\quad\quad \theta | \mu^A, \mu^B, t, F, N, \beta,\\ &\quad\quad \beta | \mu^A, \mu^B, t, F, N, \theta.\end{aligned}$$

When the matching matrix C is itself one of the parameters of interest, one can simply add, at each iteration of the algorithm, a draw from the conditional distribution

$$p(C|\mu^A \mu^B, \beta, t, F, N, \theta, x^A, x^B).$$

Details about this conditional distribution are given later in this section.

To illustrate the first updating step, notice that

$$\begin{aligned}p(\mu^A, \mu^B, t|F, \theta, \beta, x^A, x^B) &= p(\mu^A|F, \beta, x^A) p(\mu^B|F, \beta, x^B)\\ &\quad \times p(t|\mu^A, \mu^B, F, \theta, \beta, x^A, x^B).\end{aligned}$$

Moreover, by using results from Appendix B,

$$\begin{aligned}p(t|\mu^A, \mu^B, F, \theta, \beta, x^A, x^B) &= p(t|\mu^A, \mu^B, F)\\ &= \prod_{j=1}^{k} \left[\frac{\binom{f_j^A}{t_j}\binom{F_j - f_j^A}{f_j^B - t_j}}{\binom{F_j}{f_j^B}} \right]\end{aligned} \tag{4.1}$$

Thus, conditionally on all the other quantities, $t_1, \ldots, t_k$ are independent hypergeometric random variables. Then one should separately draw μ^A and μ^B from $p(\mu^A|F, \beta, x^A)$ and $p(\mu^B|F, \beta, x^B)$ and t from (4.1). However, the direct simulation of μ^A and μ^B is not straightforward. To see why, let $F_{\mu|\mu_1,\ldots,\mu_l}$ be the population count for the category assumed by μ after eliminating, from the population, l units with categories $\mu_1, \ldots, \mu_l$. Then

$$\begin{aligned}p(\mu^S|F, \beta, x^S) &\propto p(\mu^S|F) p(x^S|\mu^S, \beta)\\ &\propto \prod_{s=1}^{n^S} F_{\mu_s^S|\mu_1^S,\ldots,\mu_{s-1}^S} \prod_{i=1}^{k} \left[\beta_i I_{\{\mu_s^{S_i} = x_s^{S_i}\}} + (1 - \beta_i) \frac{1}{k_i} \right]\end{aligned}$$

for $S = A, B$ and the direct simulation from the above distributions can be computationally hard. To circumvent the difficulty of directly simulating the entire joint

distribution $p(\mu^S|F,\beta,x^S)$, note that we can easily draw the full conditionals

$$p(\mu_s^S|\mu_{-s}^S,\beta,x^S) \propto F_{\mu_s^S|\mu_1^S,\ldots,\mu_{s-1}^S,\mu_{s+1}^S,\ldots,\mu_{n^S}^S} \times \prod_{i=1}^{k}\left[\beta_i I_{\{\mu_s^{S_i}=x_s^{S_i}\}} + (1-\beta_i)\frac{1}{k_i}\right] \tag{4.2}$$

for $s=1,\ldots,n^S$ and $S=A,B$. By simulating μ^A and μ^B from (4.2) following a Gibbs type updating and t by its true conditional distribution, we do not produce an exact draw from the conditional distribution of (μ^A,μ^B,t). However, the latter is exactly the stationary distribution associated with the proposed step. This strategy can then be justified as an example of "Metropolis within Gibbs." Moreover, note that, in order to improve the mixing of the chain, for each iteration of the algorithm we can repeat more simulation cycles from the conditional distributions (4.2) in order to approximately generate, at each iteration, a random draw from the true conditional of (μ^A,μ^B,t).

A standard Gibbs updating is possible for the second step. Consider the full conditional distribution of the vector F; using the results in Appendix A and after some algebra,

$$\begin{aligned} p(F|\mu^A,\mu^B,t,\theta,\beta,x^A,x^B) &\propto p(\mu^A,\mu^B|F,t)p(t|F)p(F|\theta) \\ &\propto \prod_{j=1}^{k}\frac{F_j!}{(F_j-f_j^B-f_j^A+t_j)!}\frac{\theta_j^{F_j}}{F_j!}\frac{\Gamma(N-g+1)}{\binom{N}{n^A}\binom{N}{n^B}} \\ &\propto (N-n^A-n^B+T)!\prod_{j=1}^{k}\frac{\theta_j^{F_j-f_j^B-f_j^A+t_j}}{(F_j-f_j^A-f_j^B+t_j)!} \\ &\quad\times\frac{\Gamma(N-g+1)}{(N-n^A-n^B+T)!\binom{N}{n^A}\binom{N}{n^B}}. \end{aligned}$$

Then, random draws from the above distribution can easily be obtained by first simulating N from

$$p(N|T) \propto \frac{\Gamma(N-g+1)}{(N-n^A-n^B+T)!\binom{N}{n^A}\binom{N}{n^B}} \propto \frac{\binom{n^A}{T}\binom{N-n^A}{n^B-T}}{\binom{N}{n^B}}\frac{\Gamma(N-g+1)}{N!}. \tag{4.3}$$

Subsequently, conditionally on N, one can draw $v_1,\ldots,v_k$ from a multinomial distribution with parameters $\theta_1,\ldots,\theta_k$ and size $N-n^A-n^B+T$, and then set $F_j=v_j+f_j^A+f_j^B-t_j$.

Incidentally, we notice that the posterior distribution (4.3) plays a crucial role also when the sample sizes n^A and n^B are assumed to be random and T is known. In fact, in this case, the vector $[T,n^A-T,n^B-T,N-n^A-n^B+T]$ follows a multinomial distribution with parameters N and $(p^Ap^B,p^A(1-p^B),p^B(1-$

$p^A), (1-p^A)(1-p^B))$, where p^A and p^B represent the unknown capture probabilities in the two sampling occasions; see, for example, Bishop, Fienberg and Holland (1975). It follows that, for (p^A, p^B) unknown, inference for N can be drawn either by using the complete model [i.e., by introducing a prior for (p^A, p^B) and then getting the marginal posterior distribution $p(N|n^A, n^B, T)$] or, in a slightly approximate way, by eliminating (p^A, p^B) via a conditional argument [i.e., by using the conditional likelihood $p(T|n^A, n^B, N)$]. These two approaches typically produce very similar conclusions. In the former case, when assuming a uniform prior for (p^A, p^B), the marginal posterior of N is given by the expression for $p(N|T)$ in (4.3) multiplied by $(N+1)^{-2}$. In the latter case, inference is based only on (4.3). The complete multinomial likelihood can obviously be used within our approach by simply adding other Gibbs steps for (p^A, p^B). However, as in the case with known T, we do not expect to see substantial differences, and in the rest of the paper we will consider n^A and n^B as fixed.

The updating of θ can be done in a standard way since

$$p(\theta|\mu^A, \mu^B, t, F, N, \beta, x^A, x^B) \propto p(\theta)p(F|\theta)$$

and the independent Dirichlet distributions characterizing $p(\theta)$ are conjugate to $p(F|\theta)$; see O'Hagan and Forster (2004). Finally, note that the conditional posterior density for β_i is proportional to

$$\big(\beta_i + (1-\beta_i)/k_i\big)^{\tilde{n}_i^{AB}} (1-\beta_i)^{n^A+n^B-\tilde{n}_i^{AB}},$$

where $\tilde{n}_i^{AB}$ is the total number of sample units where the observed value and the true value coincide for the ith key variable. One can easily see that the posterior distribution of $\eta_i = \beta_i + (1-\beta_i)/k_i$, conditionally on all the other variables, is Beta$(\tilde{n}_i^{AB}+1, n^A+n^B-\tilde{n}_i^{AB}+1)$ truncated on the set $(k_i^{-1}, 1)$. Then we draw η_i from its Beta distribution and set $\beta_i = (k_i\eta_i - 1)/(k_i - 1)$, for $i = 1, \ldots, k$.

4.1. *Matching matrix simulation.* In order to specify the conditional distribution of C given all other quantities involved in the model, we introduce the sets $A_j = \{a : \mu_a^A = v_j\}$ and $B_j = \{b : \mu_b^B = v_j\}$. In words, A_j is the set of units in sample A whose true value belongs to category v_j; these sets depend on μ^A and μ^B. Let C_j be the block of the matrix C corresponding to the rows in A_j and the columns in B_j. Conditional on the true values, μ^A and μ^B, $C_{ab} = 0$ for each couple such that $\mu_a^A \neq \mu_b^B$; then, outside the blocks $C_1, \ldots, C_k$, the elements of C will be equal to 0. Thus,

$$p(C|\mu^A, \mu^B, t, F, \theta, x^A, x^B) = \prod_{j=1}^{k} p(C_j|t_j, f_j^A, f_j^B),$$

where $p(C_j|t_j, f_j^A, f_j^B)$ is the discrete uniform distribution over the set of all possible configurations for the block C_j with exactly t_j matches,

$$p(C_j|t_j, f_j^A, f_j^B) = \left[t_j! \binom{f_j^A}{t_j} \binom{f_j^B}{t_j}\right]^{-1}.$$

Note that, by conditioning on the drawn values of the key variables, we automatically create a blocking method able to limit the number of candidate matches. Blocking strategies are very popular in record linkage literature. They basically consist of a partition into homogeneous groups of all the possible comparisons among records in order to reduce the computational burden; see, for example, Newcombe (1967) or Winkler (2004). Within our approach the homogenous groups of records are identified at each step of the algorithm by the block matrices C_j's.

4.2. *Matching matrix estimation via MCMC algorithm.* Now we describe inferential strategies for producing a "point estimate" in a record linkage analysis. The usual output of an MCMC based analysis is a sample of approximately independent "observations," simulated from the posterior distribution. This sample can be used to obtain a representation of the uncertainty about the parameters of interest, mainly the matrix C or N. In addition, record linkage procedures are often the first stage of a more complex statistical analysis: they represent the crucial step of creating a suitable data set to be used afterward. In terms of statistical theory, this is equivalent to producing a point estimate of C, from which we select the "declared" matches. Classical inference methods usually provide plug-in estimates, based on theories developed in Fellegi and Sunter (1969) and Jaro (1989). First, the previously defined parameters m and u are estimated and then a sequence of statistical tests is performed in order to decide whether each pair $(a, b) \in A \times B$ can be declared a match or not. The power of multiple tests is calibrated in order to obtain a specific level of the False Match Rate (FMR), that is, the ratio between the number of false matches and the total number of declared matches. Note that the FMR is exactly equivalent to the well-known False Discovery Rate [Benjamini and Hochberg (1995)], very popular in multiple comparison applications (wavelets theory, microarray analysis, etc.) Furthermore, currently used record linkage procedures must complete the statistical data analysis with a reallocation procedure which eliminates inconsistencies among the results of different tests [see Jaro (1989) and the problem posed by Larsen (1999), paragraph 3.3].

The Bayesian way of facing a record linkage problem is different in spirit, and suggests interesting issues, both from a practical and a methodological perspective. Although in a formal Bayesian analysis one should select the point estimate as the one minimizing the posterior expected loss, it is common practice, in applications, to use the posterior mean or, sometimes, the posterior median. Of course, these solutions do not appear reasonable in a record linkage context: the marginal posterior mean of each single element of the matrix C will be a number between 0 and 1, which does not help much in deciding whether the pair (a, b) is a match or not. The use of the posterior median is even more complicated in multivariate discrete settings. Thus, a formal decision theoretic approach seems necessary: let $G = \{G_{ab}\} \in \mathcal{G}, a = 1, \ldots, n^A$ and $b = 1, \ldots, n^B$, a generic matrix of size $n^A \times n^B$, with the same characteristics as C, such that it represents our "action."

Here $\mathcal{G}$ represents the set of all possible actions. Also, let $L(\cdot,\cdot)$ be a loss function defined as $L : \mathcal{G} \times \mathcal{C} \to R^+$ where $\mathcal{C}$ is the set of all possible matching matrices. Our goal is to select, for a given loss function L, the optimal decision G^*, the one which minimizes the posterior expected loss

$$G^* = \underset{G \in \mathcal{G}}{\operatorname{argmin}}\, W(G)$$

where $W(G) = E[L(C, G)|x^A, x^B]$. In what follows we will consider some specific loss functions:

(1) Quadratic Loss

$$L_q(C, G) = \sum_a \sum_b (C_{ab} - G_{ab})2.$$

Since the elements of C and G are either 0 or 1, L_q is equivalent to the L_1 loss: $L_1(C, G) = \sum_a \sum_b |C_{ab} - G_{ab}|$.

(2) False Match Rate

$$L_{\mathrm{FMR}}(C, G) = \begin{cases} 0, & \text{if } \sum_a \sum_b G_{ab} = 0, \\ \dfrac{\sum_a \sum_b G_{ab} I(C_{ab} = 0)}{\sum_a \sum_b G_{ab}}, & \text{otherwise.} \end{cases}$$

L_{FMR} translates, in terms of decision theory, the classical use of the False Match Rate as a measure of performance of the record linkage analysis.

(3) Absolute number of errors

$$L_{\mathrm{ABS}}(C, G) = \sum_a \sum_b [G_{ab} I(C_{ab} = 0) + (1 - G_{ab}) I(C_{ab} = 1)].$$

The following theorem provides the optimal solution for the above mentioned losses.

THEOREM 4.1.

(A) *Under losses L_q and L_{ABS}, the optimal Bayesian solution is given by the matrix G^*, defined as*

$$G^*_{ab} = \begin{cases} 1, & \text{if } p(C_{ab} = 1|x^A, x^B) > \frac{1}{2}, \\ 0, & \text{otherwise,} \end{cases}$$

$$a = 1, \ldots, n^A; b = 1, \ldots, n^B.$$

(B) *Under loss L_{FMR}, the optimal solution is a matrix consisting of all zeros.*

PROOF. First, notice that $I(C_{ab} = 1) = C_{ab}$ and $I(C_{ab} = 0) = 1 - C_{ab}$.
(A): Since

$$L_q(C, G) = \sum_a \sum_b [C_{ab} + G_{ab} - 2C_{ab}G_{ab}],$$

the problem is equivalent to the maximization of the posterior expected value of

$$L_q(C,G) = 2\sum_a \sum_b G_{ab}\left[C_{ab} - \frac{1}{2}\right].$$

With the loss L_{ABS}, simple calculations lead to

$$\begin{aligned} L_{\mathrm{ABS}}(C,G) &= \sum_a \sum_b [G_{ab}(1-C_{ab}) + (1-G_{ab})C_{ab}] \\ &= \sum_a \sum_b [G_{ab} - 2G_{ab}C_{ab} + C_{ab}]. \end{aligned}$$

The minimization of the posterior expected loss of L_{ABS} is equivalent to the maximization of the quantity

$$L_q(C,G) = 2\sum_a \sum_b G_{ab}\left[C_{ab} - \frac{1}{2}\right].$$

Then the quantities L_q and L_{ABS} are identical and it will be sufficient to find the optimal solution for L_q. We need to maximize

$$\begin{aligned} W_q(G) &= 2E\left(\sum_a \sum_b G_{ab}\left[C_{ab} - \frac{1}{2}\right] \Big| x^A, x^B\right) \\ &= 2\sum_a \sum_b G_{ab}\left[p(C_{ab} = 1 | x^A, x^B) - \frac{1}{2}\right]. \end{aligned}$$

The last expression shows that the value that maximizes $W_q(G)$ is obtained by setting $G_{ab} = 1$ if and only if the correspondent coefficient is positive, that is, when $p(C_{ab} = 1|x^A, x^B) > \frac{1}{2}$.

(B): When L_{FMR} is used, it is easy to see that FMR is minimized by adopting the conservative behavior of not declaring any match! In this case, in fact, the posterior expected loss is always zero, independently of the posterior distribution. Then the optimal solution is given by $G^*_{ab} = 0$, for all (a,b). □

It is important to stress that all the optimal solutions derived in Theorem 4.1 are based on the marginal posterior probabilities of being a match for the various pairs (a,b). This is a consequence of the fact that the above loss functions are additive and they basically "sum" over all the losses due to the single mismatches.

Part B of Theorem 4.1 is also important. It says that, from a decision theoretic perspective, the FMR is not a valid measure of performance, because it only controls one type of error. Every reasonable loss function should also take into account a measure of the number of undiscovered matches [Genovese and Wasserman (2003)]. In this sense, a reasonable loss function for record linkage may be given by the *Global Error Rate*

$$L_{\mathrm{TOT}}(C,G) = L_{\mathrm{FMR}}(C,G) + \frac{\sum_a \sum_b (1-G_{ab}) I_{C_{ab}=1}(C_{ab})}{\sum_a \sum_b (1-G_{ab})}.$$

The loss L_{TOT} is actually able to capture errors due to missing true matches. However, the improvement is more theoretical than practical: in fact, the denominator of the second factor is so much larger than the denominator of L_{FMR} that the results obtained using L_{TOT} should not be practically different from those derived under loss L_{FMR}.

5. Illustrative application. We illustrate our approach in detail with the real data set already used in the Introduction. The two files consist of $n^A = 34$ records from a single block of the last Italian census population survey and $n^B = 45$ records from the same block relative to the post enumeration survey; more details can be found in Alleva, Fortini and Tancredi (2007). Records in both files refer to foreign residents only, which typically represent an example of an elusive population. For each file, we take three key variables: X_1 represents the first two consonants of the family name with 339 observed categories (considering all blocks), X_2 represents the gender and X_3 is the education level, with 17 categories. The total number of entries in V is $k = 11{,}526$. The data and the programs [written in C and R, R Development Core Team (2009)] that have been used for this application are available in the supplementary material [Tancredi and Liseo (2011)]. In practice, real applications may have more key variables, more blocks and larger sample sizes. However, focusing on a small example allows us to illustrate better some details of our methodology compared to the existing approaches.

The hyperparameter g appearing in the prior distribution $p(N)$ has been set equal to 2 in order to have a proper prior. The Dirichlet distributions for θ are chosen so that, at the superpopulation level, X_1 is independent of (X_2, X_3). We also assume that all the Dirichlet distributions are uniform in their supports.

We have used the algorithm described in Section 4 to generate a single Markov chain of length 100,000. See the supplementary material for a graphical representation of some of the simulation traces. Figure 2 shows the posterior distributions of the following quantities: (a) the number of matches T, (b) the total population size N, (c) the measurement error parameter vector β_i, $(i = 1, 2, 3)$, (d) the probability of selecting a male within the block at the superpopulation level, $\theta_{\cdot 1}$.. In panel (d), we also show the posterior distributions of $\theta_{\cdot 1}$. obtained by considering the two files separately, assuming a uniform prior and independence among the units. Notice that the posterior density of $\theta_{\cdot 1}$. can be graphically interpreted as an average of the two posteriors one would have obtained from the analysis of each single data set.

The posterior estimated quantiles of level $(0.05, 0.5, 0.975)$ for T are $(26, 28, 31)$. The same posterior summaries for N are $(49, 55, 65)$. Marginal posterior probabilities of being a match, $p(C_{ab} = 1|x^A, x^B)$, are graphically displayed in panel (a) of Figure 3, where the cases have been sorted in order to have the most probable matches on the diagonal. There are only 34 pairs of records (out of 1530) such that $p(C_{ab} = 1|x^A, x^B)$ is larger than 0.1. The estimated matching matrix, using the quadratic loss function outlined in Section 4, is given by the 27 matches

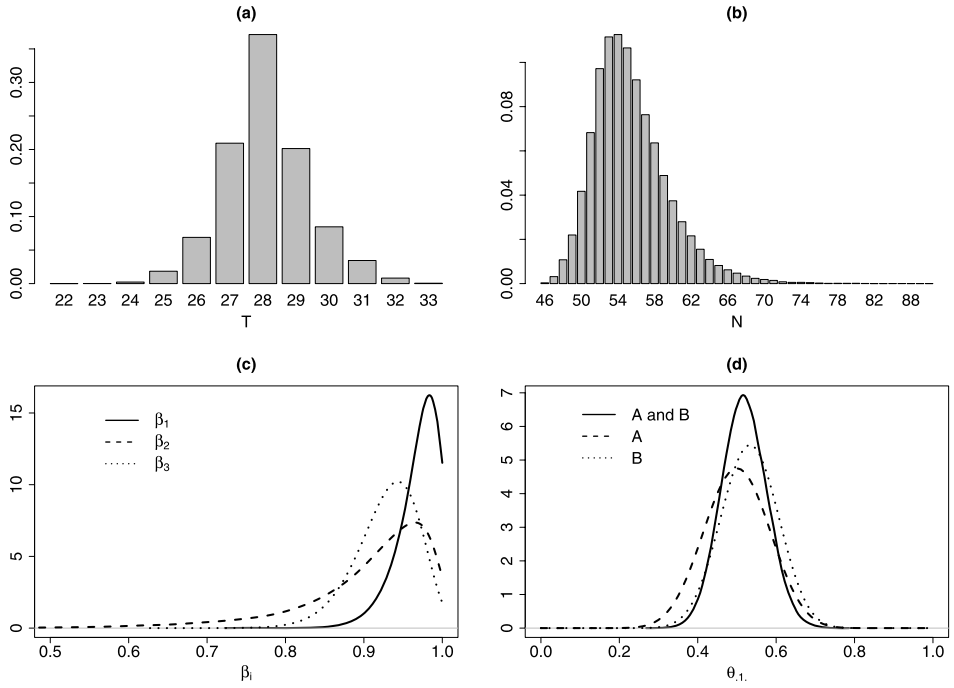

FIG. 2. *Posterior distributions of the number of matches T, the population size N, the parameter vector β_i $(i = 1, 2, 3)$ and $\theta_{.1}$..*

visible on the diagonal. Notice that inference about C is quite robust with respect to the choice of the hyperparameter g: when $g = 1$ we obtained exactly the same estimated matching matrix, while setting $g = 3$ would produce one more match.

We now compare our results with other possible approaches based on the comparison vectors y_{ab} whose frequency distribution is given in Table 2. As a first alternative we consider a slight modification of the Bayesian approach proposed by Larsen (2005) where y_{ab} is marginally distributed as (2.1) and the matching matrix C satisfies the constraints $\sum_a C_{ab} \leq 1$ and $\sum_b C_{ab} \leq 1$. We use uniform priors for m and u. Unlike Larsen (2005), we have assumed, for the matching matrix C, the same prior distribution used in our approach. We will call this model the "Jaro constrained model." The posterior distribution for the parameters (m, u, C, N) can easily be simulated by using Gibbs steps for $[m|u, C, N]$, $[u|m, C, N]$ and $[N|u, m, C]$. To update the matching matrix C, we use the Metropolis–Hastings step proposed by Green and Mardia (2006). Figure 4 reports the posterior distributions of the parameters $p = T/(n^A \cdot n^B)$, m and u. The posterior quantiles of level $(0.05, 0.5, 0.975)$ for T are estimated as $(23, 27, 31)$. The same posterior summaries for N are $(49, 57, 72)$. The marginal posterior probabilities of being a match, $p(C_{ab} = 1|y_{11}, \ldots, y_{n^A,n^B})$, are graphically displayed in panel (b) of Figure 3. Also in this case we have exactly 34 pairs of records such that

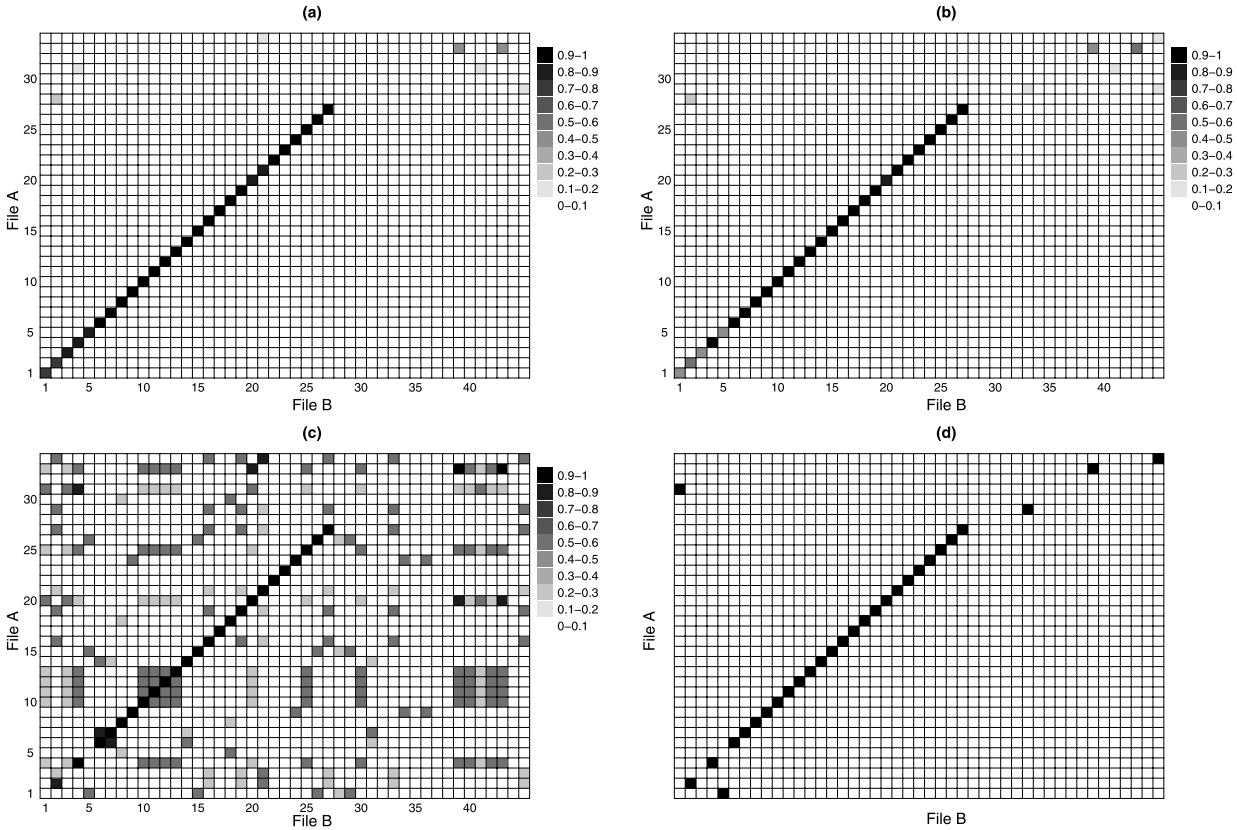

FIG. 3. *Matching estimation. Panel* (a) *shows the posterior probabilities* $p(C_{ab} = 1|x^A, x^B)$ *under the new model. Panel* (b) *shows the posterior probabilities* $p(C_{a,b} = 1|y_{11}, \ldots, y_{n^A,n^B})$ *under the Jaro constrained model. Panels* (c) *and* (d) *show the posterior probabilities* $p(C_{ab} = 1|y_{ab})$ *and the estimated matching matrix using the classical approach. Values of the posterior probabilities are indicated by the shading scale at the right of each panel.*

TABLE 2
Results of the classic approach

y_{ab}	Frequency	$p((a,b) \in M \mid y_{ab})$	λ
(0, 0, 0)	659	0.00	0.01
(1, 0, 0)	20	0.01	0.14
(0, 1, 0)	601	0.00	0.04
(1, 1, 0)	13	0.05	0.58
(0, 0, 1)	78	0.23	3.43
(1, 0, 1)	8	0.80	45.20
(0, 1, 1)	126	0.56	14.81
(1, 1, 1)	25	0.94	194.97

The first two columns give the distribution of the comparison vector. The last two columns report the estimated quantities (2.3) and (2.2).

$p(C_{ab} = 1 | y_{11}, \ldots, y_{n^A n^B})$ is larger than 0.1, but the matching matrix obtained with the quadratic loss provides 25 matches. In general, our proposed model and the Jaro constrained model provide similar estimates, although the latter seems to produce slightly more uncertainty as shown by the larger interval estimates for both T and N.

Finally, we show the results obtained by considering model (2.1) without row or column constraints on the matching matrix C. Maximum likelihood estimates and posterior densities are reported in Figure 4. The matching step is performed by considering the posterior matching probabilities (2.3) or the likelihood ratios (2.2). In Table 2 we report these quantities obtained with a simple plug-in of the maximum likelihood estimates of the parameters. The posterior probabilities $p((a,b) \in M | y_{ab})$ are also displayed graphically in panel (c) of Figure 3. In this case there are 237 pairs with a posterior probability $p((a,b) \in M | y_{ab})$ greater than 0.1. The higher number of *potential* matches is almost certainly due to the fact that, in this approach, because of the independence assumption among comparison vectors and the absence of constraints on the C matrix, the marginal matching probabilities only depend on the information retrieved from the single comparison and not, as in the previous models, on the information provided by the entire data set. To rule out multiple matches, following Jaro (1989), we maximize the function

$$\sum_{a=1}^{n^A} \sum_{b=1}^{n^B} z_{ab} \log \frac{\prod_{i=1}^{k} (\hat{m}^{y^i_{ab}} (1-\hat{m})^{1-y^i_{ab}})}{\prod_{h=1}^{k} (\hat{u}^{y^i_{ab}} (1-\hat{u})^{1-y^i_{ab}})} \tag{5.1}$$

subject to the constraints $\sum_{a=1}^{\nu_A} z_{ab} \leq 1 \ \forall b$, $\sum_{b=1}^{\nu_B} z_{ab} \leq 1 \ \forall a$ and $z_{ab} \in \{0,1\}$ $\forall (a,b)$. The final answer produces 29 matches displayed in panel (d) of Figure 3. From Table 1 one can see that, by setting $T = 29$ in the hypergeometric likelihood $\binom{n^A}{T}\binom{N-n^A}{n^B-T}/\binom{N}{n^B}$ and using the prior $p(N) \propto 1/N^2$, one gets a 95% credible inter-

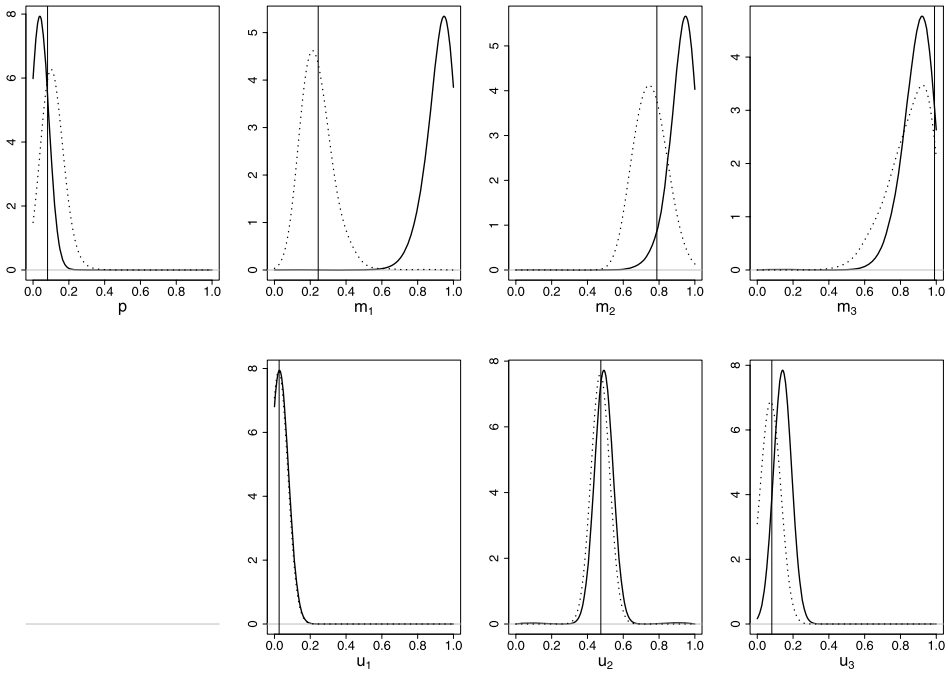

FIG. 4. *Posterior distributions for the parameters of model* (2.1) *with the constraints on the matching matrix C* (*solid line*) *and without* (*dotted line*). *For the latter case the constraint* $p < 1/2$ *has been used to guarantee identifiability and the vertical lines indicate maximum likelihood estimates*.

val for N equal to $[50, 60]$, which is a subset of the intervals obtained using our approach or the Jaro constrained model.

6. Multiple block application. In this section we illustrate the results obtained with a more realistic exercise involving a multiple block scenario. In particular, we repeated the analysis described in the previous section for each census enumeration area (census block) also selected for the post enumeration survey and including at least one foreign person during the census survey. This way we obtained a list with 337 pairs of data sets for a total of 3675 records taken on foreign people during the 2001 census population survey and 3404 analogous records originating from the parallel post enumeration survey. The block sizes vary from a minimum of one individual on at least one occasion to a maximum with 280 and 311 individuals on the two occasions. Note that the total number of blocks selected for the post enumeration survey is 1098, corresponding to 0.31% of the total number of the Italian census enumeration areas.

For each pair of data sets we performed a record linkage analysis in order to estimate the total number of foreign people N_l living in the lth census block, for $l = 1, \ldots, 337$. In addition to the three key variables considered in the single block

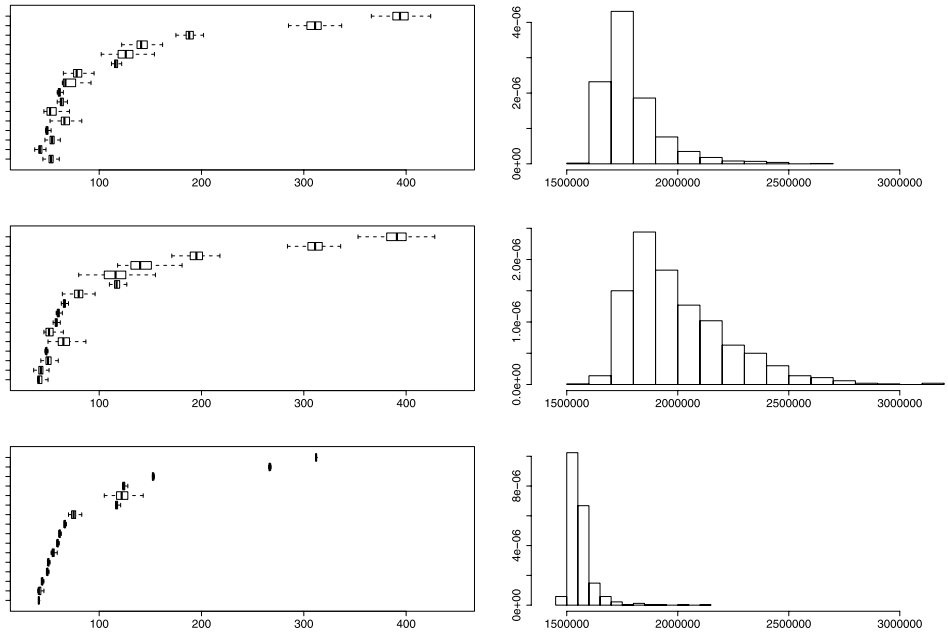

FIG. 5. Left panels: *box-plots from the posterior distribution of the foreign population size for blocks with at least* 25 *matches*. Right panels: *posterior distribution of the foreign population size in Italy at the end of* 2001. Upper panels: *new model*. Central panels: *Jaro constrained model*. Lower panels: *Jaro unconstrained hybrid approach*.

analysis outlined before, we also considered the age (coded into 10 categories). At the superpopulation level in our hierarchical model we assumed the surname to be independent of gender, education level and age. The probability vector for the surname categories is assumed, as before, to be uniform in its support. For the $340 = 2 \times 17 \times 10$ joint probabilities of the other three key variables we set the Dirichlet hyperparameters all equal to $1/340$ in order to avoid marginal distributions that are too concentrated.

In the upper left panel of Figure 5 we show, for each block with approximately at least 25 matches, the box-plot for the posterior distribution of N_l given by our approach. For each pair of data sets we also implemented the other two approaches described throughout the illustrative example, namely, the Jaro constrained model and the hybrid strategy obtained by estimating the matching matrix via the classical approach and then plugging in the estimated match number in the posterior distribution of population size. The box-plots for the posterior distribution of N_l obtained with these two approaches are shown, respectively, in the central and lower left panels of Figure 5. Note that the posterior distributions for N_l provided by the Jaro constrained model give point estimates similar to those obtained with our approach but with slightly wider credibility intervals. Instead, the pattern shown by the hybrid strategy is quite different. In particular, when compared to the other

approaches, it shows a remarkable under-estimation of the block sizes. In fact, the maximization of the function (5.1) leads to an over-estimation of the true match number. However, introducing a false match rate correction as in Belin and Rubin (1995) would reduce the distance from the other approaches. Nevertheless, there is a clear message that ignoring the matching uncertainty would give a false impression of accuracy for the estimates.

The same conclusions are emphasized when we aim at estimating the quantity $N = [\sum_{l=1}^{337} N_l]/0.0031$ which can be seen as a rough approximation for the size of the foreign population in Italy at the end of 2001. The histograms shown in the right panels of Figure 5 have been obtained by summing the draws from the posterior distributions of N_l for each block with at least 2 records in both the surveys. In fact, smaller blocks tend to produce quite diffuse posterior distributions, making the MCMC inference difficult without introducing a more concentrated prior. To overcome this problem, the population size for the smaller blocks has been fixed equal to $\hat{N}_l = (n_l^A + 1)(n_l^B + 1)/(\hat{T}_l + 1)$ with $\hat{T}_l$ estimated by the classical approach. In particular, one can notice that accounting for matching uncertainty with the Jaro constrained model (central right panel) produces both a larger estimate and larger uncertainty with respect to our approach (upper right panel).

7. Simulation studies. We now evaluate our hierarchical model via a simulation study. Artificial data are often used to evaluate record linkage techniques, especially in computer science literature; see, for example, Christen (2005) and Christen and Pudjijono (2009). Here, we consider three main different simulation scenarios generating, at the superpopulation level, three and six independent key variables (scenarios 1 and 3) and three dependent key variables (scenario 2). Common features across different simulations are as follows:

- the population size, fixed at $N = 100$.
- the sample size; we always assume $n^A = n^B$ equal to 70, 80, 90.
- the measurement error parameters β's: their value has been fixed at (0.85, 0.90, 0.95).

In the first two scenarios, the three key variables assume, respectively, 64, 16 and 4 categories, leading to a contingency table with 4096 entries. In the independence case the means of the population frequencies F_j have been set equal to $\theta_j = \theta_{j_1,j_2,j_3} = \prod_{i=1}^{3} 3b_{j_i}$ where $b_{j_i} \propto j_i$ with $j_i = 1, \ldots, k_i$ for $i = 1, 2, 3$. Under the dependence model we set $\theta_j = \theta_{j_1,j_2,j_3} = b_{j_3} b_{j_2|j_3} b_{j_1|j_3}$ where $= b_{j_3} \propto j_3$, $b_{j_2|j_3} \propto j_2^{j_3}$ and $b_{j_1|j_3} \propto j_1^{1/j_3}$. Finally, in the third scenario, the 6 key variables assume, respectively, 32, 16, 4, 4, 2 and 2 categories, leading to a contingency table with 32,768 cells with $\theta_j = \theta_{j_1,\ldots,j_6} = \prod_{i=1}^{6} b_{j_i}$ where $b_{j_i} \propto j_i$.

For each combination of model parameters, we have generated 100 pairs of data sets. Each pair of data sets has been analyzed using our hierarchical model with 45,000 iterations of the MCMC algorithm and 5000 iterations discarded for burn-in. For each pair of data sets we also implemented the Jaro constrained model and

the hybrid strategy described in the previous sections. Mixing and convergence rates were satisfactory based on the examination of trace plots.

In Table 3 we focus on the inference for N. For each of the three approaches and for each group of 100 pairs of data sets we report the average values of the

TABLE 3

Simulation study for evaluating the posterior mean $E(N)$ and the 95% credible interval under the new model (M_1), the Jaro constrained model (M_2) and the Jaro unconstrained hybrid approach (M_3)

		$E(N)$			Coverage			Length		
β_i	n^S	M_1	M_2	M_3	M_1	M_2	M_3	M_1	M_2	M_3
Scenario 1: independence with 3 key variables										
0.95	90	101 (0.40)	102 (0.38)	103 (0.85)	0.92	0.96	0.21	16 (0.48)	18 (0.50)	6 (0.45)
	80	99 (0.55)	103 (0.68)	98 (0.93)	0.96	0.96	0.30	24 (0.65)	28 (0.81)	9 (0.53)
	70	96 (0.76)	101 (0.85)	92 (1.18)	0.95	0.97	0.29	35 (0.81)	38 (0.80)	12 (0.78)
0.90	90	103 (0.55)	107 (0.84)	103 (1.10)	0.97	0.91	0.20	26 (0.83)	32 (1.30)	6 (0.60)
	80	100 (0.78)	110 (1.02)	98 (1.09)	0.96	0.91	0.22	36 (1.06)	48 (1.35)	9 (0.65)
	70	96 (1.06)	108 (1.35)	90 (1.14)	0.93	0.93	0.20	50 (1.52)	62 (1.57)	11 (0.71)
0.85	90	104 (0.72)	115 (1.52)	102 (1.22)	0.99	0.79	0.18	37 (1.32)	56 (2.39)	6 (0.71)
	80	100 (0.87)	116 (1.53)	93 (0.94)	0.99	0.93	0.17	51 (1.49)	75 (2.61)	6 (0.46)
	70	97 (1.37)	120 (2.10)	86 (1.24)	0.99	0.94	0.11	69 (2.87)	101 (4.30)	9 (0.84)
Scenario 2: dependence with 3 key variables										
0.95	90	101 (0.40)	103 (0.50)	99 (0.31)	0.90	0.90	0.26	16 (0.49)	19 (0.56)	4 (0.15)
	80	99 (0.59)	103 (0.61)	93 (0.52)	0.96	0.96	0.24	26 (0.58)	29 (0.67)	6 (0.29)
	70	95 (0.74)	101 (0.81)	85 (0.59)	0.94	0.97	0.07	36 (1.02)	39 (0.97)	8 (0.37)
0.90	90	102 (0.49)	111 (0.74)	99 (0.45)	0.96	0.82	0.23	27 (0.74)	37 (1.05)	4 (0.22)
	80	100 (0.74)	111 (1.07)	92 (0.84)	0.94	0.90	0.15	38 (1.09)	50 (1.49)	6 (0.49)
	70	93 (0.96)	107 (1.26)	85 (1.16)	0.92	0.93	0.08	45 (1.41)	60 (1.79)	8 (0.84)
0.85	90	104 (0.66)	120 (1.27)	99 (0.75)	0.98	0.74	0.21	38 (1.21)	58 (2.16)	4 (0.40)
	80	100 (1.05)	122 (1.89)	90 (0.71)	0.98	0.82	0.14	51 (1.92)	77 (2.82)	5 (0.41)
	70	95 (1.21)	123 (1.84)	82 (0.74)	0.98	0.94	0.05	63 (2.30)	104 (3.69)	6 (0.48)
Scenario 3: independence with 6 key variables										
0.95	90	101 (0.27)	101 (0.29)	102 (0.44)	0.83	0.83	0.27	10 (0.25)	11 (0.28)	5 (0.23)
	80	101 (0.45)	102 (0.51)	100 (0.64)	0.93	0.95	0.59	19 (0.51)	20 (0.47)	10 (0.37)
	70	99 (0.72)	101 (0.74)	95 (0.79)	0.94	0.94	0.59	28 (0.65)	30 (0.67)	14 (0.72)
0.90	90	103 (0.40)	104 (0.54)	102 (0.56)	0.90	0.87	0.26	17 (0.44)	20 (0.65)	5 (0.31)
	80	100 (0.66)	104 (0.83)	95 (0.63)	0.98	0.93	0.39	26 (0.80)	32 (1.00)	8 (0.35)
	70	98 (0.94)	104 (0.11)	89 (0.10)	0.95	0.91	0.24	40 (1.11)	45 (1.24)	11 (0.58)
0.85	90	105 (0.65)	111 (0.99)	101 (0.59)	0.88	0.82	0.17	29 (0.94)	40 (1.61)	5 (0.29)
	80	100 (0.95)	112 (1.65)	93 (0.85)	0.93	0.85	0.15	38 (1.34)	54 (2.42)	6 (0.47)
	70	100 (1.22)	118 (1.85)	85 (0.91)	0.98	0.89	0.14	58 (1.96)	81 (3.10)	8 (0.56)

Each cell reports a mean obtained with 100 pairs of data sets drawn from our model with $N = 100$, $n^A = n^B = n^S$ and $\beta_1 = \cdots = \beta_i = \cdots = \beta_h$. Standard errors are in parentheses.

posterior mean of N, the estimated coverage of the 95% credibility intervals and their mean length. Estimated standard errors are also given in parentheses. Note that in our approach the average value of the posterior mean for N is, in almost every experimental condition, the closest to the true value $N = 100$ and the one with the smallest standard error. However, the reduced bias of our approach was to be expected because the simulation generating process is exactly part of our model, while the other approaches present several misspecification elements. Note also that the Jaro constrained model and the hybrid approach have different behaviours, the former overestimating N and the latter underestimating it. This is the same trend already observed in the multiple block application.

The performance of the alternative approaches does not improve when considering the interval estimates. In fact, with few exceptions, our approach produces the interval estimates with a coverage level closest to the nominal one. The hybrid approach, as expected, has dramatically low coverage level since it does not account for matching uncertainty. The Jaro constrained model always produces interval estimates wider than those provided by our model, partly because it only retrieves from the data the marginal information given by the comparisons.

It is also interesting to note the behavior of the estimates with respect to the information carried by the data. When the sample sizes or the number of key variables increase, the uncertainty about N reduces with all three methods. In addition, both our model and the Jaro constrained model show a decrement in uncertainty as the measurement error level decreases, that is, when the β_i's approach 1.

In Table 4 we report the results regarding the estimation of the matching matrix C. In particular, for each method we show the average value of the False Match Rates defined by

$$\text{FMR1} = \frac{\sum_{ab} \hat{C}_{ab}(1 - C_{ab})}{\sum_{ab} \hat{C}_{ab}} \quad \text{and} \quad \text{FMR2} = \frac{\sum_{ab} C_{ab}(1 - \hat{C}_{ab})}{\sum_{ab} C_{ab}},$$

where $\hat{C}$ is the point estimate obtained using the quadratic loss. The results of the comparisons among different methods would depend upon which type of FMR is used. In particular, under the FMR1 criterion, the better performance is established by the Jaro constrained model, followed by our approach and by the hybrid approach. However, one should recall that the Jaro constrained model tends to overestimate N and, consequently, it leads to a potential under-estimation of $T = \sum_{ab} C_{ab}$. This way, the FMR1 criterion would prefer the Jaro constrained approach, because of its "conservative behavior." From our perspective this is another argument in favor of the inadequacy of FMR1 as a single measure of performance of record linkage procedures. Finally, note that, when using the FMR2 criterion, the hybrid approach quite often shows the better performance with our model producing a lower rate than the Jaro constrained model under the independence assumption.

TABLE 4

Simulation study for evaluating the false match rates under the new model (columns M_1), the Jaro constrained model (columns M_2) and the Jaro unconstrained hybrid approach (columns M_3)

		FMR1			FMR2		
β_i	n^S	M_1	M_2	M_3	M_1	M_2	M_3
		Scenario 1: independence with 3 key variables					
0.95	90	0.063 (0.004)	0.052 (0.003)	0.101 (0.004)	0.129 (0.004)	0.111 (0.005)	0.126 (0.005)
	80	0.074 (0.004)	0.058 (0.004)	0.146 (0.007)	0.147 (0.006)	0.148 (0.006)	0.125 (0.006)
	70	0.085 (0.005)	0.073 (0.005)	0.203 (0.008)	0.165 (0.006)	0.168 (0.006)	0.130 (0.006)
0.90	90	0.095 (0.004)	0.088 (0.004)	0.185 (0.006)	0.240 (0.007)	0.244 (0.008)	0.216 (0.006)
	80	0.100 (0.005)	0.090 (0.005)	0.244 (0.007)	0.274 (0.007)	0.286 (0.007)	0.212 (0.006)
	70	0.123 (0.006)	0.110 (0.006)	0.319 (0.009)	0.293 (0.007)	0.309 (0.008)	0.227 (0.007)
0.85	90	0.130 (0.006)	0.122 (0.006)	0.307 (0.007)	0.401 (0.007)	0.401 (0.008)	0.316 (0.007)
	80	0.131 (0.005)	0.122 (0.006)	0.373 (0.008)	0.423 (0.008)	0.429 (0.007)	0.320 (0.009)
	70	0.160 (0.007)	0.144 (0.008)	0.420 (0.009)	0.447 (0.010)	0.457 (0.010)	0.322 (0.007)
		Scenario 2: dependence with 3 key variables					
0.95	90	0.065 (0.003)	0.054 (0.003)	0.138 (0.004)	0.137 (0.005)	0.126 (0.005)	0.123 (0.005)
	80	0.075 (0.004)	0.067 (0.005)	0.205 (0.005)	0.173 (0.006)	0.152 (0.005)	0.144 (0.006)
	70	0.083 (0.005)	0.087 (0.005)	0.278 (0.006)	0.184 (0.006)	0.170 (0.006)	0.144 (0.006)
0.90	90	0.093 (0.004)	0.091 (0.004)	0.234 (0.007)	0.270 (0.006)	0.268 (0.006)	0.227 (0.006)
	80	0.108 (0.005)	0.108 (0.005)	0.295 (0.008)	0.289 (0.007)	0.283 (0.008)	0.228 (0.008)
	70	0.117 (0.006)	0.123 (0.007)	0.353 (0.008)	0.308 (0.007)	0.298 (0.008)	0.234 (0.006)
0.85	90	0.128 (0.005)	0.145 (0.005)	0.339 (0.007)	0.426 (0.007)	0.405 (0.007)	0.333 (0.007)
	80	0.141 (0.007)	0.139 (0.007)	0.398 (0.007)	0.438 (0.007)	0.422 (0.008)	0.334 (0.007)
	70	0.145 (0.007)	0.149 (0.008)	0.454 (0.007)	0.463 (0.010)	0.454 (0.010)	0.331 (0.008)
		Scenario 3: independence with 6 key variables					
0.95	90	0.034 (0.002)	0.030 (0.002)	0.043 (0.003)	0.054 (0.003)	0.046 (0.003)	0.058 (0.003)
	80	0.043 (0.003)	0.040 (0.003)	0.079 (0.005)	0.065 (0.003)	0.065 (0.003)	0.070 (0.004)
	70	0.068 (0.004)	0.061 (0.004)	0.134 (0.006)	0.078 (0.005)	0.083 (0.005)	0.073 (0.004)
0.90	90	0.071 (0.003)	0.063 (0.003)	0.115 (0.004)	0.143 (0.005)	0.140 (0.005)	0.124 (0.005)
	80	0.089 (0.004)	0.078 (0.003)	0.175 (0.006)	0.158 (0.005)	0.168 (0.005)	0.132 (0.005)
	70	0.104 (0.005)	0.097 (0.005)	0.244 (0.007)	0.190 (0.006)	0.204 (0.007)	0.149 (0.006)
0.85	90	0.126 (0.005)	0.108 (0.004)	0.232 (0.006)	0.287 (0.006)	0.297 (0.007)	0.235 (0.006)
	80	0.142 (0.006)	0.129 (0.006)	0.297 (0.008)	0.308 (0.009)	0.330 (0.010)	0.241 (0.008)
	70	0.151 (0.007)	0.131 (0.006)	0.373 (0.008)	0.342 (0.009)	0.380 (0.011)	0.256 (0.008)

Each cell reports a mean obtained with 100 pairs of data sets drawn from our model with $N = 100$, $n^A = n^B = n^S$ and $\beta_1 = \cdots = \beta_i = \cdots = \beta_h$. Standard errors are in parentheses.

8. Discussion. Record linkage techniques pose several interesting problems both from the methodological and the computational viewpoint. From a methodological perspective, the definition itself of the statistical framework within which

comparisons among records should be performed is still under debate: in this paper we have proposed a novel Bayesian methodology.

While it is definitely true that the result of a statistical analysis produced by an official organism must be objective (or — at least — it should be perceived as such by the users), it is also undeniable [see Fienberg (2011)] that Bayesian ideas and techniques can play an important role in official statistics, especially when important prior (or extra-experimental) information about the variables of interest exists and cannot be adequately exploited in a classical inference framework. In addition, even when prior information is lacking, a Bayesian analysis may be necessary simply because a classical approach cannot provide answers without introducing strong assumptions, not easily testable. In these situations a Bayesian analysis allows us, at least, to perform a sensitivity analysis, with the aim of quantifying the influence of the assumptions on inferences.

From a computational perspective record linkage problems become formidable as soon as the sizes of the files are large. The intensive simulation methods required by any Bayesian approach for a matching problem make the computational problems in real applications even more crucial. One of the most popular solutions, valid also for our approach, is to perform the record linkage only between those records which show the same values on some blocking variables which are assumed to be recorded without errors. In addition, parallel computations for separated blocks may reduce the computing time in a significant way.

The proposed model is built up on the actually observed categorical variables drawn from a finite population and no reduction of the available information, for example, by using Boolean comparison vectors, takes place. We also stress that prior information, provided by experts or by previous surveys, can be introduced naturally into the record linkage process via the superpopulation model, for example, by giving specific association patterns between the key variables. Another important benefit is the acknowledgment and incorporation of the matching process uncertainty in estimating the population size as well as other population parameters. At the same time, the information available about the population parameters and their uncertainty are accounted for in the record linkage.

Throughout the paper we have made some specific assumptions, such as the fixed sample sizes or the uniform distribution for the misspecified record fields and their conditional independence given the true values. Anyway, we are confident that our framework may provide a basis for several extensions with more general assumptions. In particular, some of the capture–recapture models used for a closed population [see, e.g., Wolter (1986) and Fienberg, Johnson and Junker (1999) or Erosheva, Fienberg and Joutard (2007) and Manrique-Vallier and Fienberg (2008) for more advanced proposals] could be incorporated as sampling models for the sample sizes and the number of recaptures. Multiple recaptures could be handled following Ruffieux and Green (2009), where a method for aligning multiple unlabeled configurations has been proposed. By assuming an exchangeable prior for $\beta_1, \ldots, \beta_h$, we may also remove the assumption of conditional independence for

the measurement error among record fields. In addition, different measurement error probabilities across files may be considered. Note also that the model has been developed so that each block is separately evaluated. However, following Larsen (2005), we could allow a "borrowing of strength" effect across the blocks by introducing some extra layers in our prior modeling. Some of these extensions will be the object of future research. A similar approach for handling multivariate normal data is discussed in Liseo and Tancredi (2009).

An important aspect of record linkage procedures which we have not addressed here is that of the nonrandomness of the samples, for example, in applications using administrative lists provided by register offices. This issue has some consequences in every modeling approach to record linkage; however, discussion about these problems is beyond the scope of this paper. In any case, we believe that the idea of a Bayesian superpopulation model generating the lists might be useful in this context too.

Finally, note that the computer science literature on record linkage (also known as data matching or entity resolution) has developed, in recent times, some impressive algorithms based on machine learning and graph-based matching. Some relevant papers are Bhattacharya and Getoor (2007) and Kalashnikov and Mehrotra (2006) and it would be interesting to compare these or similar approaches with the statistical models presented in this paper.

APPENDIX A

The sampling models (3.1) and (3.2) can be obtained as the marginal distribution of $p(\mu^A, \mu^B, C, t|F) = p(\mu^A, \mu^B|C, t, F)p(C, t|F)$. First, we average out C, so $p(\mu^A, \mu^B|t, F) = \sum_C p(\mu^A, \mu^B, C|t, F)$. In this sum we only need to consider those matrices C with exactly t_j matches in the block $\{(a, b) : \mu_a^A = \mu_b^B = v_j\}$ for $j = 1, \ldots, k$. The total number of such matrices is $\prod_{j=1}^k \binom{f_j^A}{t_j}\binom{f_j^B}{t_j} t_j!$. Also,

$$
\begin{aligned}
p(\mu^A, \mu^B|t, F) &= \sum_C p(\mu^A, \mu^B, C|t, F) = \sum_C p(\mu^A, \mu^B|C, t, F)p(C|t, F) \\
&= p(\mu^A, \mu^B|C, t, F)p(C|t, F) \prod_{j=1}^k \binom{f_j^A}{t_j}\binom{f_j^B}{t_j} t_j! \\
&= \frac{\prod_{j=1}^k \binom{F_j - t_j}{f_j^A - t_j, f_j^B - t_j, F_j - f_j^A - f_j^B + t_j}}{\binom{N-T}{n^A - T, n^B - T, N - n^A - n^B + T}} \frac{\prod_{j=1}^k f_j^A f_j^B}{n^A! n^B!} \\
&= \frac{1}{\binom{n^A}{f_1^A, \ldots, f_k^A}} \frac{1}{\binom{n^B}{f_1^B, \ldots, f_k^B}} \frac{\prod_{j=1}^k \binom{F_j - f_j^A}{f_j^B - t_j}\binom{F_j - t_j}{f_j^A - t_j}}{\binom{N - n^A}{n^B - T}\binom{N-T}{n^A - T}}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
p(\mu^A, \mu^B|F) &= \sum_t p(\mu^A, \mu^B, t|F) = \sum_t p(\mu^A, \mu^B|t, F)p(t|F) \\
&= \frac{1}{\binom{n^A}{f_1^A,\ldots,f_k^A}} \frac{1}{\binom{n^B}{f_1^B,\ldots,f_k^B}} \sum_t \frac{\prod_{j=1}^k \binom{F_j-f_j^A}{f_j^B-t_j}\binom{F_j-t_j}{f_j^A-t_j}}{\binom{N-n^A}{n^B-T}\binom{N-T}{n^A-T}} \frac{\binom{F_j}{t_j}}{\binom{N}{T}} \frac{\binom{n^A}{T}\binom{N-n^A}{n^B-T}}{\binom{N}{n^B}} \\
&= \frac{1}{\binom{n^A}{f_1^A,\ldots,f_k^A}} \frac{1}{\binom{n^B}{f_1^B,\ldots,f_k^B}} \frac{1}{\binom{N}{n^A}\binom{N}{n^B}} \\
&\quad \times \sum_t \prod_{j=1}^k \frac{F_j!}{(F_j - f_j^A - f_j^B + t_j)!(f_j^A - t_j)!(f_j^B - t_j)!t_j!} \\
&= \frac{1}{\binom{n^A}{f_1^A,\ldots,f_k^A}} \frac{1}{\binom{n^B}{f_1^B,\ldots,f_k^B}} \frac{\prod_{j=1}^k \binom{F_j}{f_j^A}\binom{F_j}{f_j^B}}{\binom{N}{n^A}\binom{N}{n^B}}.
\end{aligned}
$$

APPENDIX B

The derivation of the full conditional distribution of $t|F, \mu^A, \mu^B, \theta, \beta, x^A, x^B$:

$$
\begin{aligned}
&p(t|F, \mu^A, \mu^B, \theta, \beta, x^A, x^B) \\
&\quad \propto p(\mu^A, \mu^B|F, t)p(t|F) \\
&\quad \propto \frac{\prod_{j=1}^k \binom{F_j-t_j}{f_j^A-t_j}}{\binom{N-T}{n^A-T}} \frac{\prod_{j=1}^k \binom{F_j-f_j^A}{f_j^B-t_j}}{\binom{N-n^A}{n^B-T}} \frac{\prod_{j=1}^k \binom{F_j}{t_j}}{\binom{N}{T}} \frac{\binom{n^A}{T}\binom{N-n^A}{n^B-T}}{\binom{N}{n^B}} \\
&\quad \propto \frac{\prod_{j=1}^k \left(\frac{F_j!}{t_j!(f_j^A-t_j)!(f_j^B-t_j)!(F_j-f_j^A-f_j^B+t_j)!}\right)}{\binom{N}{n^A}\binom{N}{n^B}} \\
&\quad \propto \prod_{j=1}^k \frac{\binom{f_j^A}{t_j}\binom{F_j-f_j^A}{f_j^B-t_j}}{\binom{F_j}{f_j^B}}.
\end{aligned}
$$

Acknowledgments. We would like to thank the referees, the Associate Editor and the Editor (Professor Stephen E. Fienberg) for their very helpful comments and suggestions that substantially improved a previous version of this manuscript.

SUPPLEMENTARY MATERIAL

Data files and codes (DOI: 10.1214/10-AOAS447SUPP; .zip). Included in the supplementary material there are the following files: exampleA.dat,

exampleB.dat and exampleV.dat contain the data used in Section 5. The files B.Cat.matching.example.R, example.R, functions.r, gibbs.c contain the codes. The file supplementary_figure.pdf shows the trace plots for the application described in Section 5.

REFERENCES

ALBERT, P. S. and DOOD, L. E. (2004). A cautionary note on robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics* **60** 427–435.

ALLEVA, G., FORTINI, M. and TANCREDI, A. (2007). The control of non-sampling errors on linked data: An application on population census. In *Proceedings of the* 2007 *Intermediate Conference. Risk and Prediction. Venice*.

ARMSTRONG, J. and MAYDA, J. E. (1993). Model-based estimation of record linkage error rates. *Survey Methodology* **19** 137–147.

BELIN, T. R. and RUBIN, D. B. (1995). A method for calibrating false—match rates in record linkage. *J. Amer. Statist. Assoc.* **90** 694–707.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300.

BHATTACHARYA, I. and GETOOR, L. (2007). Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data* (*TKDD*) **1** Article No. 5.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Parctice*. MIT Press, Cambridge.

CHRISTEN, P. (2005). Probabilistic data generation for deduplication and data linkage. In *IDEAL'*05. *LNCS* **3578** 109–116. Springer, Berlin.

CHRISTEN, P. and PUDJIJONO, A. (2009). Accurate synthetic generation of realistic personal information. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (*PAKDD'*09). *LNAI* **5476** 507–514. Springer, Berlin.

COPAS, J. B. and HILTON, F. J. (1990). Record linkage: Statistical models for matching computer records. *J. Roy. Statist. Soc. Ser. A* **153** 287–320.

DARROCH, J. N. (1958). The multiple capture census I. Estimation of a closed population. *Biometrika* **45** 343–358.

DEGROOT, M. H. and GOEL, P. K. (1980). Estimation of the correlation coefficient from a broken random sample. *Ann. Statist.* **8** 264–278.

DO, K. A., MUELLER, P. and TANG, F. (2005). A Bayesian mixture model for differential gene expression. *J. Roy. Statist. Soc. Ser. C.* **54** 627–644.

ERICSON, W. A. (1969). Subjective Bayesian models in sampling finite populations. *J. Roy. Statist. Soc. Ser. B* **31** 195–224.

EROSHEVA, E., FIENBERG, S. and JOUTARD, C. (2007). Describing disability through individual level mixture models for multivariate binary data. *Ann. Appl. Statist.* **1** 502–537.

FELLEGI, I. P. and SUNTER, A. B. (1969). A theory of record linkage. *J. Amer. Statist. Assoc.* **64** 1183–1210.

FIENBERG, S. E. (2011). Bayesian models and methods in public policy and government settings. *Statist. Sci.* To appear.

FIENBERG, S. E., JOHNSON, M. S. and JUNKER, B. W. (1999). Classical multilevel and Bayesian approaches to population size estimation using multiple list, part 3. *J. Roy. Statist. Soc. Ser. A* **162** 383–405.

FIENBERG, S. E. and MANRIQUE-VALLIER, D. (2009). Integrated methodology for multiple systems estimation and record linkage using a missing data formulation. *Advances in Statistical Analysis* **93** 49–60.

FORSTER, J. J. and WEBB, E. L. (2007). Bayesian disclosure risk assesment: Predicting small frequencies in contingency table. *J. Roy. Statist. Soc. Ser. C* **56** 551–570.

FORTINI, M., LISEO, B., NUCCITELLI, A. and SCANU, M. (2001). On Bayesian record linkage. *Research in Official Statistics* **4** 185–198.

FORTINI, M., LISEO, B., NUCCITELLI, A. and SCANU, M. (2002). Modelling issues in record linkage: A Bayesian perspective. In *Proceedings of the Section on Survey Research Methods* 1008–1013. Amer. Statist. Assoc., Alexandria, VA.

GENOVESE, C. and WASSERMAN, L. (2003). Bayesian and frequentist multiple testing (with discussion). In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 145–161. Oxford Univ. Press, New York. MR2003171

GREEN, P. J. and MARDIA, K. V. (2006). Bayesian alignment using hierarchical models, with applicationin protein bioinformatics. *Biometrika* **93** 235–254.

HERZOG, T. N., SCHEUREN, F. J. and WINKLER, W. E. (2007). *Data Quality and Record Linkage Techniques*. Springer, New York.

HOADLEY, B. (1969). The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *J. Amer. Statist. Assoc.* **64** 216–229.

JARO, M. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *J. Amer. Statist. Assoc.* **84** 414–420.

JUDSON, D. H. (2007). Information integration for constructing social statistics: History, theory and ideas towards a reserach program. *J. Roy. Statist. Soc. Ser. A* **170** 483–501.

KALASHNIKOV, D. V. and MEHROTRA, S. (2006). Domain-independent data cleaning via analysis of entity-relationship graph. *ACM Transactions on Database Systems* **31** 716–767.

KELLEY, P. (1986). Robustness of the Census Bureau's record linkage system. In *Proceedings of the Section on Survey Research Methods*, *American Statistical Association* 620–624.

LAHIRI, P. and LARSEN, M. D. (2005). Regression analysis with linked data. *J. Amer. Statist. Assoc.* **100** 222–230.

LARSEN, M. D. (1999). Multiple imputation analysis of records linked using mixture model. In *Proceedings of Survey Methods Section*, *Statistical Society of Canada* 65–71. Statistical Society of Canada, Ottawa.

LARSEN, M. D. (2004). Record Linkage using finite mixture models. In *Applied Bayesian Modeling and Causal Inference from Incomplete Data Perspectives* (A. Gelman and X.-L. Meng, eds.) 309–318. Wiley, New York. MR2003171

LARSEN, M. D. (2005). Advances in record linkage theory: Hierarchical Bayesian record linkage theory. In *Proceedings of the Section on Survey Research Methods* 3277–3283. Amer. Statist. Assoc., Alexandria, VA.

LARSEN, M. D. and RUBIN, D. B. (2001). Iterative automated record linkage using mixture models. *J. Amer. Statist. Assoc.* **96** 32–41.

LINDLEY, D. V. (1977). A problem in forensic science. *Biometrika* **64** 207–213.

LINK, W. A., YOSHIZAKI, J., BAILEY, L. L. and POLLOCK, K. H. (2009). Uncovering a latent multinomial: Analysis of mark–recapture data with misidentication. *Biometrics* **66** 178–185.

LISEO, B. and TANCREDI, A. (2009). Bayesian estimation of population size via linkage of multivariate Normal data sets. Working Paper 66. Dept. Methods and Models for Economics, Territory and Finance, Sapienza Univ. Rome.

MANRIQUE-VALLIER, D. and FIENBERG, S. E. (2008). Population size estimation using individual level mixture models. *Biom. J.* **50** 1051–1063.

MARIN, J. M. and ROBERT, C. P. (2007). *Bayesian Core. A Practical Approach to Computational Statistics*. Springer, New York.

MCGLINCY, M. (2004). A Bayesian record linkage methodology for multiple imputation of missing data. In *Proceedings of the Section on Survey Research Methods* 4001–4008. Amer. Statist. Assoc., Alexandria, VA.

NEWCOMBE, H. B. (1967). Record linkage: The design of efficient systems for linking records into individual and family histories. *American Journal of Human Genetics* **9** 335–359.

NEWCOMBE, H. B., KENNEDY, J. M., AXFORD, S. J. and JAMES, A. P. (1959). Automatic linkage of vital records. *Science* **130** 954–959.

NORÉN, G. N., ORRE, R. and BATE, A. (2005). A hit–miss model for duplicate detection in the WHO drug safety database. In *KDD'*05: *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* 459–468. Canada.

O'HAGAN, A. and FORSTER, J. (2004). *Kendall's Advanced Theory of Statistics*. *Volume* 2*B*. *Bayesian Inference*. Arnold, London.

PEPE, M. S. (2003). *The Statistical Evaluation of Medical Test for Classification and Prediction*. Oxford Univ. Press, London.

PEPE, M. S. and JANES, H. (2007). Insights into latent class analysis of diagnostic test performance. *Biostatistics* **8** 474–484.

PEREZ, C. J., GIRON, F. J., MARTIN, J., RUIZ, M. and ROJANO, C. (2007). Misclassified multinomial data: A Bayesian approach. *Rev. R. Acad. Cien. Serie A. Mat* **101** 71–80.

R DEVELOPMENT CORE TEAM (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

ROBERT, C. P. and CASELLA, G. (2004). *Monte Carlo Statistical Methods*, 2nd ed. Springer, New York.

RUFFIEUX, Y. and GREEN, P. J. (2009). Alignment of multiple configurations using hierachical models. *J. Comput. Graph. Statist.* **18** 756–773.

SEBER, G. A. F. (1986). A review of estimating animal abundance. *Biometrics* **42** 267–292.

SWARTZ, T., HAITOVSKY, Y., VEXLER, A. and YANG, T. (2004). Bayesian identifiability and misclassification in multinomial data. *Canad. J. Statist.* **32** 285–302.

TANCREDI, A. and LISEO, B. (2011). Supplement to "A hierarchical Bayesian approach to record linkage and population size problems." DOI: 10.1214/10-AOAS447SUPP.

WINKLER, W. E. (1993). Improved decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods* 274–279. Amer. Statist. Assoc., Alexandria, VA.

WINKLER, W. E. (1995). Matching and record linkage. In *Buisness Survey Methods* (B. G. Cox, D. A. Binder, B. N. Chinnappa, A. Christianson, M. J. Colledge and P. S. Kott, eds.) 355–384. Wiley, New York.

WINKLER, W. E. (2000). Machine learning, information retrieval and record linkage. In *Proceedings of the Section on Survey Research Methods* 20–29. Amer. Statist. Assoc., Alexandria, VA.

WINKLER, W. E. (2004). Approximate string comparator search strategies for very large administrative lists. In *Proceedings of the Section on Survey Research Methods* 4595–4602. Amer. Statist. Assoc., Alexandria, VA.

WOLTER, K. M. (1986). Some coverage error models for census data. *J. Amer. Statist. Assoc.* **81** 338–345.

WRIGHT, J. A., BAKER, R. J., SCHOFIELD, M. R., FRANTZ, A. C., BYROM, A. E. and GLEESON, D. M. (2009). Incorporating genotype uncertainty into mark-recapture-type models for estimating abundance using DNA samples. *Biometrics* **65** 833–840.

DEPARTMENT OF METHODS AND MODELS
FOR ECONOMICS, TERRITORY AND FINANCE
SAPIENZA UNIVERSITÀ DI ROMA
VIALE DEL CASTRO LAURENZIANO 9, ROMA 00161
ITALY
E-MAIL: andrea.tancredi@uniroma1.it
brunero.liseo@uniroma1.it